معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute

NADEEF: A Commodity Data Cleaning System

Data analytics, QCRI

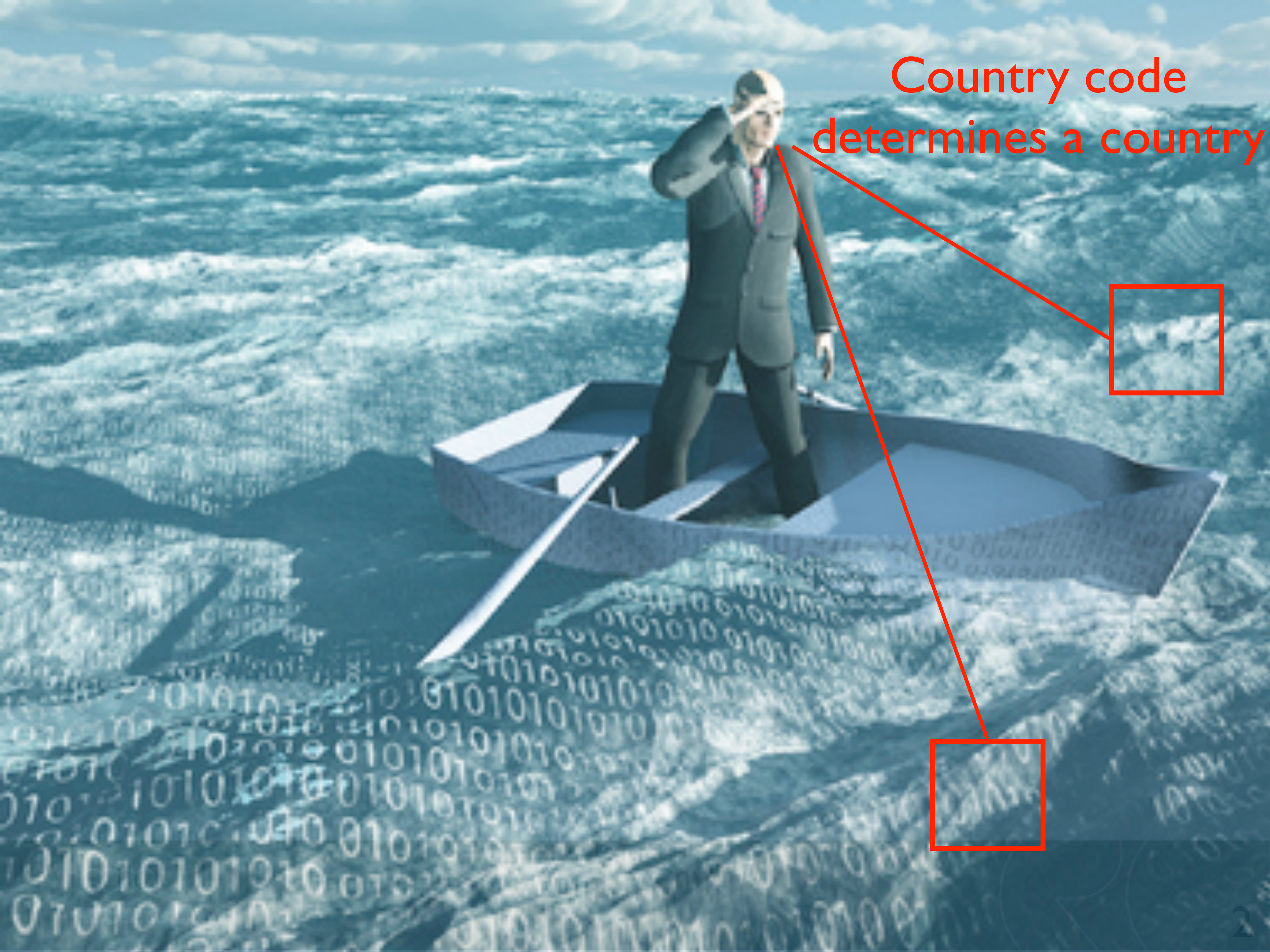
Michele Dallachiesa
University of Trento

Amr Ebaid
Purdue University

Ahmed Eldawy
University of Minnesota

Ahmed Elmagarmid Ihab F. Ilyas Mourad Ouzzani Nan Tang

Bob should be standardized to Robert



Country code determines a country

A Motivating Scenario

tran

name	street	city	CC	country	phn	when	where
David	Holywell	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
Paul	Ratcliffe	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
David	Holywell	Oxford	44	Netherlands	66700541	6am 6/05/2012	US
Paul	Market	Amsterdam	31	UK	55384922	9am 6/02/2012	Netherlands

bank

name	street	city	CC	country	tel	gd
David	Holywell	Oxford	44	UK	66700543	M
Paul	Ratcliffe	Oxford	44	UK	44944631	M

A Motivating Scenario

tran

name	street	city	CC	country	phn	when	where
David	Holywell	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
Paul	Ratcliffe	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
David	Holywell	Oxford	44	Netherlands	66700541	6am 6/05/2012	US
Paul	Market	Amsterdam	31	UK	55384922	9am 6/02/2012	Netherlands

bank

name	street	city	CC	country	tel	gd
David	Holywell	Oxford	44	UK	66700543	M
Paul	Ratcliffe	Oxford	44	UK	44944631	M

If a customer's CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands;

ETL rules (lookup table)

Extended CFDs

A Motivating Scenario

tran

name	street	city	CC	country	phn	when	where
David	Holywell	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
Paul	Ratcliffe	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
David	Holywell	Oxford	44	Netherlands	66700541	6am 6/05/2012	US
Paul	Market	Amsterdam	31	UK	55384922	9am 6/02/2012	Netherlands

bank

name	street	city	CC	country	tel	gd
David	Holywell	Oxford	44	UK	66700543	M
Paul	Ratcliffe	Oxford	44	UK	44944631	M

If the same person from different tables has different phones, the phone number from table bank is more reliable;

Editing rules
(*w.r.t.* master data)

A Motivating Scenario

tran

name	street	city	CC	country	phn	when	where
David	Holywell	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
Paul	Ratcliffe	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
David	Holywell	Oxford	44	Netherlands	66700541	6am 6/05/2012	US
Paul	Market	Amsterdam	31	UK	55384922	9am 6/02/2012	Netherlands

bank

name	street	city	CC	country	tel	gd
David	Holywell	Oxford	44	UK	66700543	M
Paul	Ratcliffe	Oxford	44	UK	44944631	M

A country code (CC) uniquely determines a country

CFDs (FDs)

A Motivating Scenario

tran

name	street	city	CC	country	phn	when	where
David	Holywell	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
Paul	Ratcliffe	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
David	Holywell	Oxford	44	Netherlands	66700541	6am 6/05/2012	US
Paul	Market	Amsterdam	31	UK	55384922	9am 6/02/2012	Netherlands

bank

name	street	city	CC	country	tel	gd
David	Holywell	Oxford	44	UK	66700543	M
Paul	Ratcliffe	Oxford	44	UK	44944631	M

If two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour, these two purchases might be a fraud.

Write a special-purpose application

A Motivating Scenario

tran

name	street	city	CC	country	phn	when	where
David	Holywell	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
Paul	Ratcliffe	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
David	Holywell	Oxford	44	Netherlands	66700541	6am 6/05/2012	US
Paul	Market	Amsterdam	31	UK	55384922	9am 6/02/2012	Netherlands

bank

name	street	city	CC	country	tel	gd
David	Holywell	Oxford	44	UK	66700543	M
Paul	Ratcliffe	Oxford	44	UK	44944631	M

If two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour, these two purchases might be a fraud.

Write a special-purpose application

Challenging to capture multiple types of rules

The User Perspective

These are our data quality rules

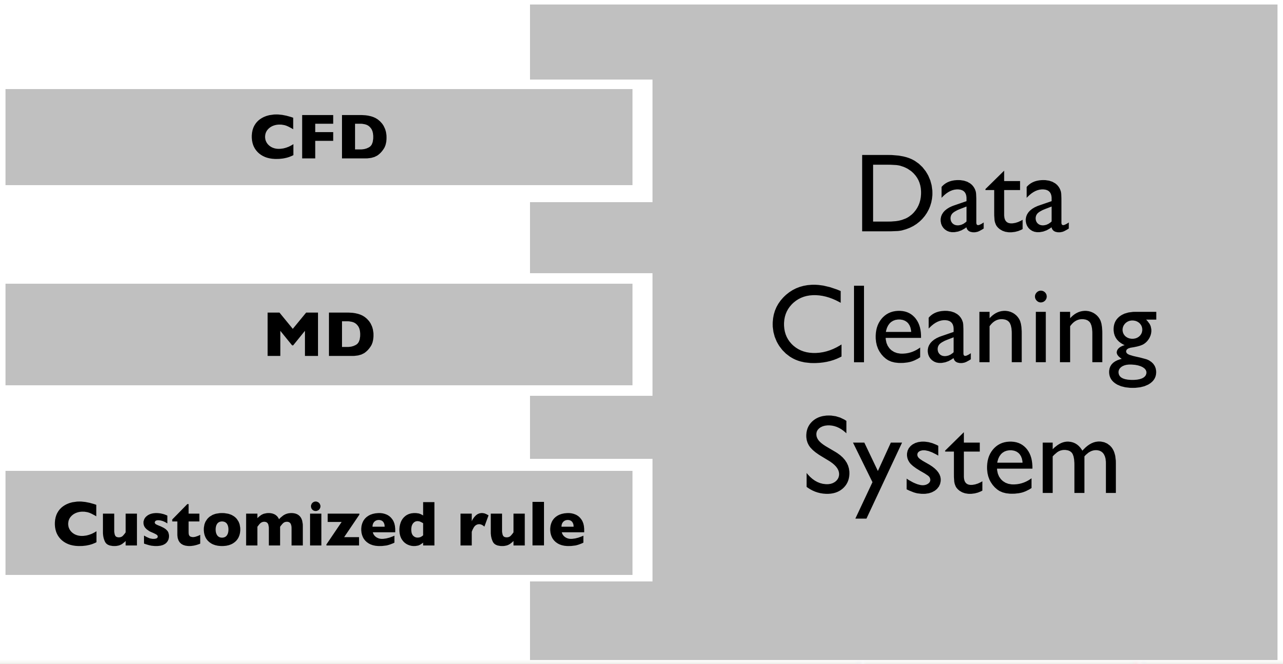
CFD

MD

Customized rule

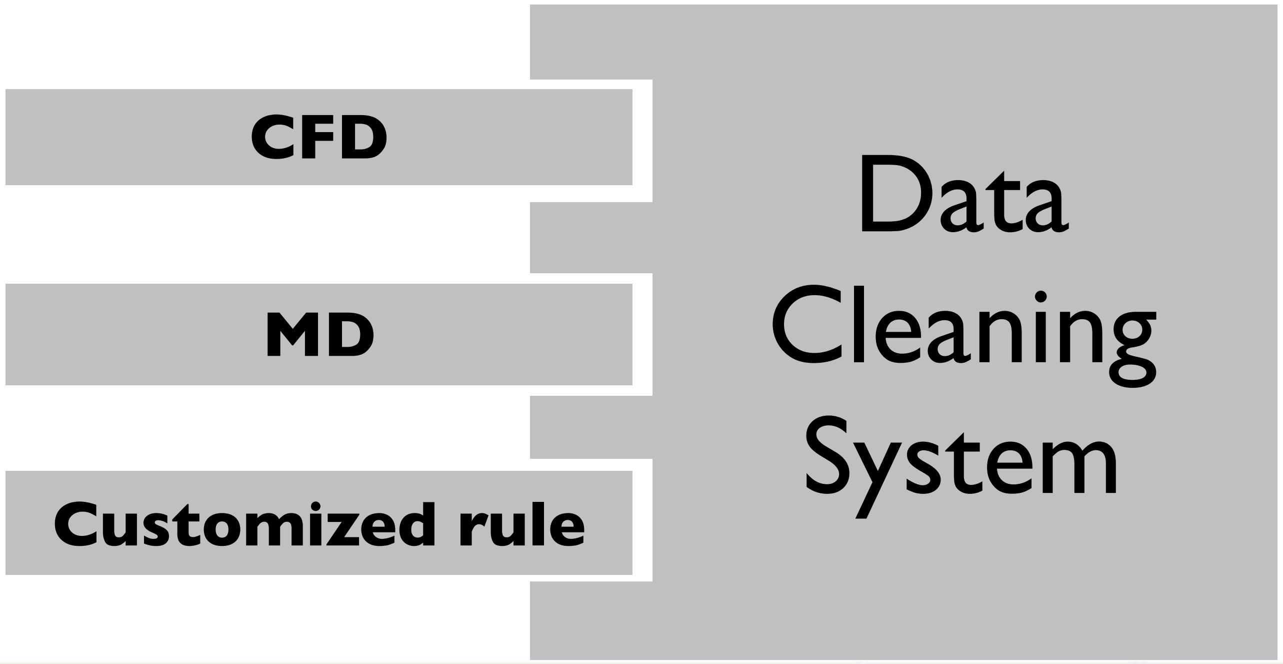
The User Perspective

These are our data quality rules



The User Perspective

These are our data quality rules



CFD

MD

Customized rule

Data Cleaning System

Easy to specify and easy to deploy

Challenges

- Heterogeneity
- Interdependency
- Deployment and extensibility
- Metadata management and user interaction

Challenges

- Heterogeneity
- Interdependency
- Deployment and extensibility
- Metadata management and user interaction

- Integrity constraints (CFDs, DCs) ETL rules, customized rules

Challenges

- Heterogeneity
- Interdependency
- Deployment and extensibility
- Metadata management and user interaction

- Integrity constraints (CFDs, DCs) ETL rules, customized rules
- Interaction of various types of rules

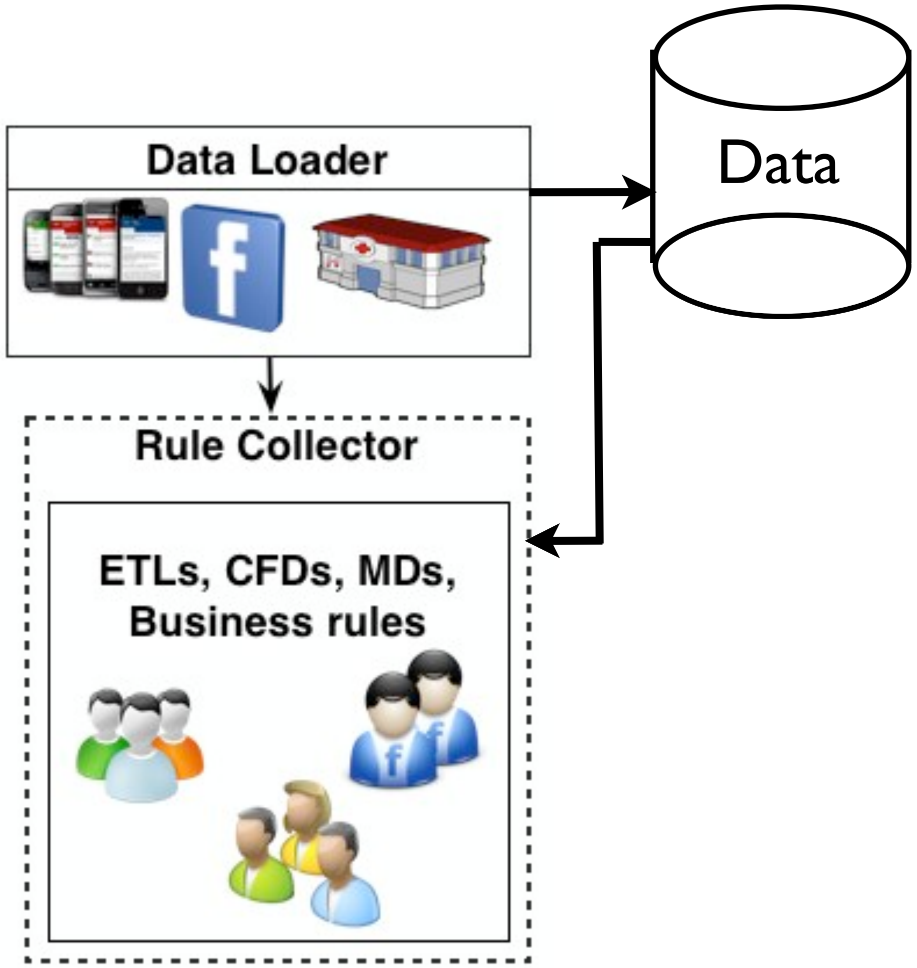
Challenges

• Heterogeneity	• Integrity constraints (CFDs, DCs) ETL rules, customized rules
• Interdependency	• Interaction of various types of rules
• Deployment and extensibility	• Download, compile and run Extend with new cleaning solutions
• Metadata management and user interaction	

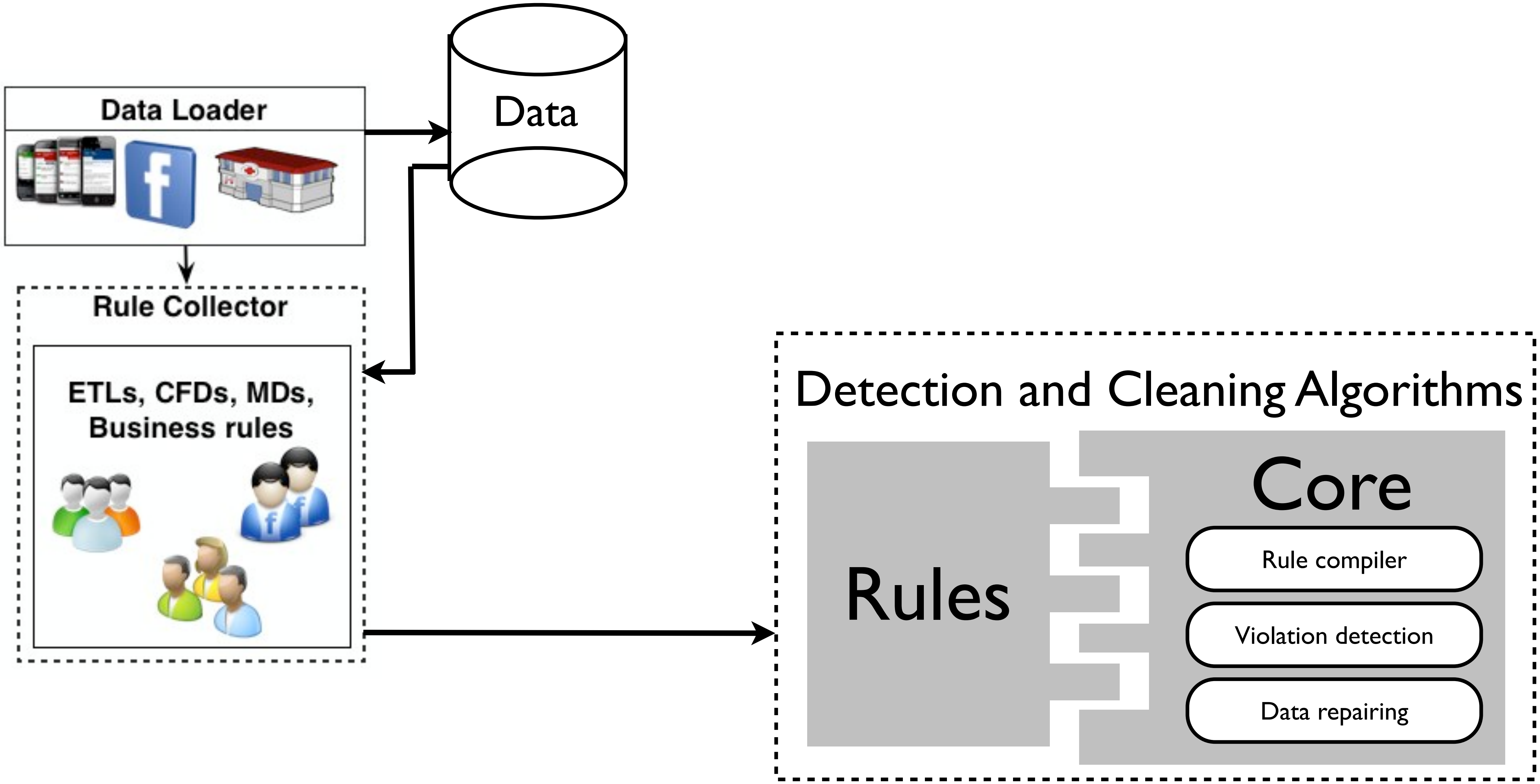
Challenges

- Heterogeneity
  - Integrity constraints (CFDs, DCs) ETL rules, customized rules
- Interdependency
  - Interaction of various types of rules
- Deployment and extensibility
  - Download, compile and run Extend with new cleaning solutions
- Metadata management and user interaction
  - Dashboard and metadata profiling

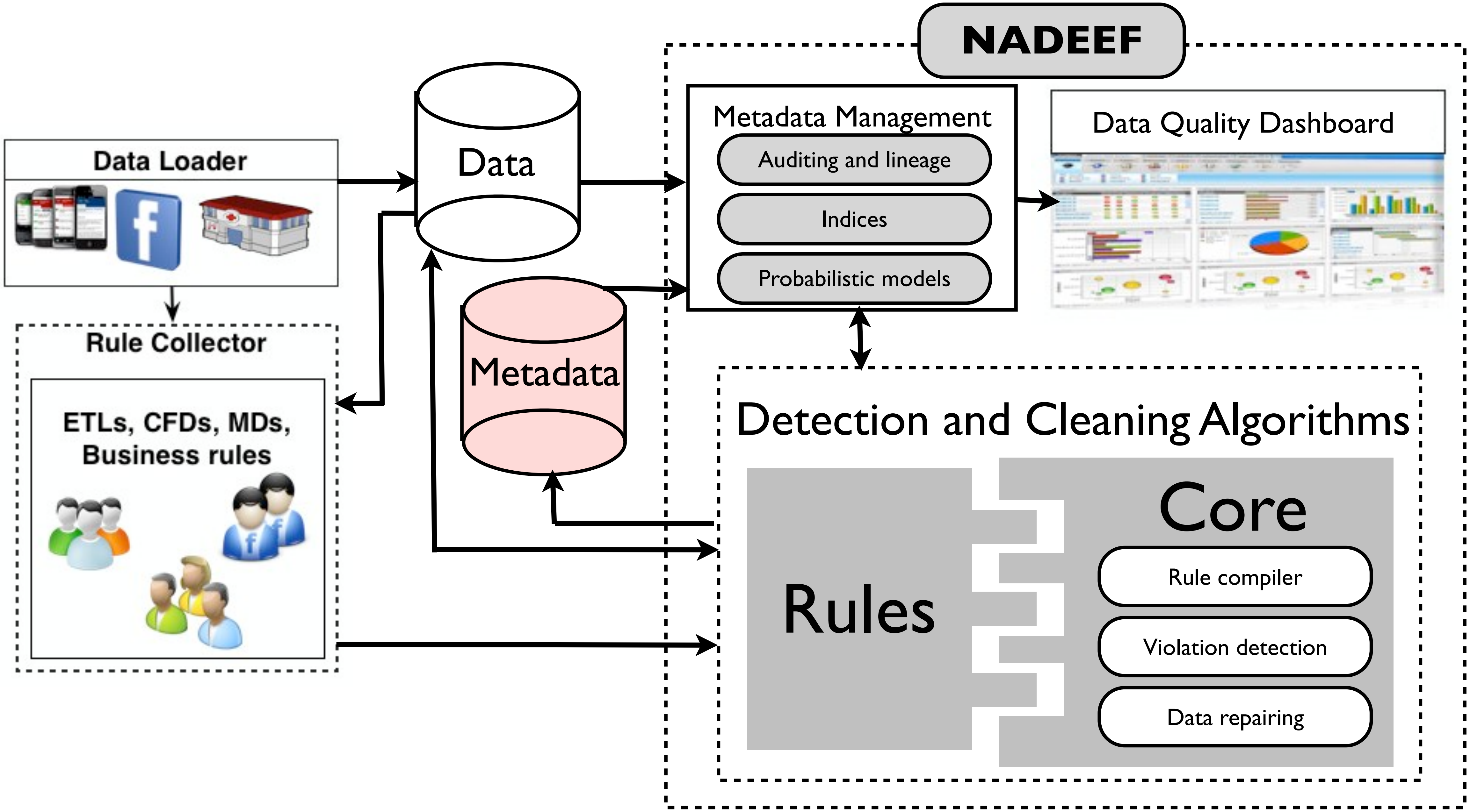
NADEEF Architecture



NADEEF Architecture



NADEEF Architecture



NADEEF

Data Loader

Data

Rule Collector

ETLs, CFDs, MDs, Business rules

Metadata

Metadata Management

Auditing and lineage

Indices

Probabilistic models

Data Quality Dashboard

Detection and Cleaning Algorithms

Rules

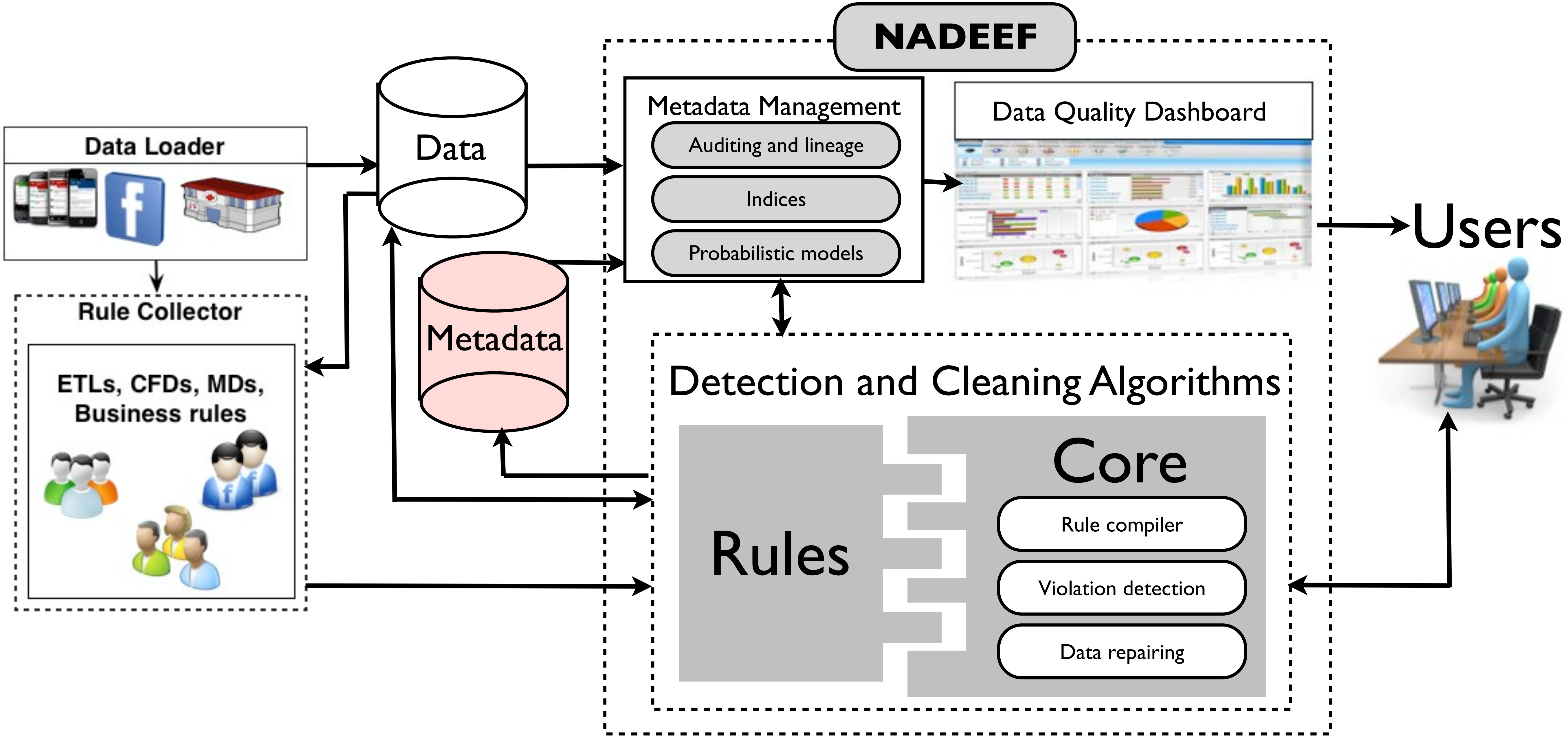
Core

Rule compiler

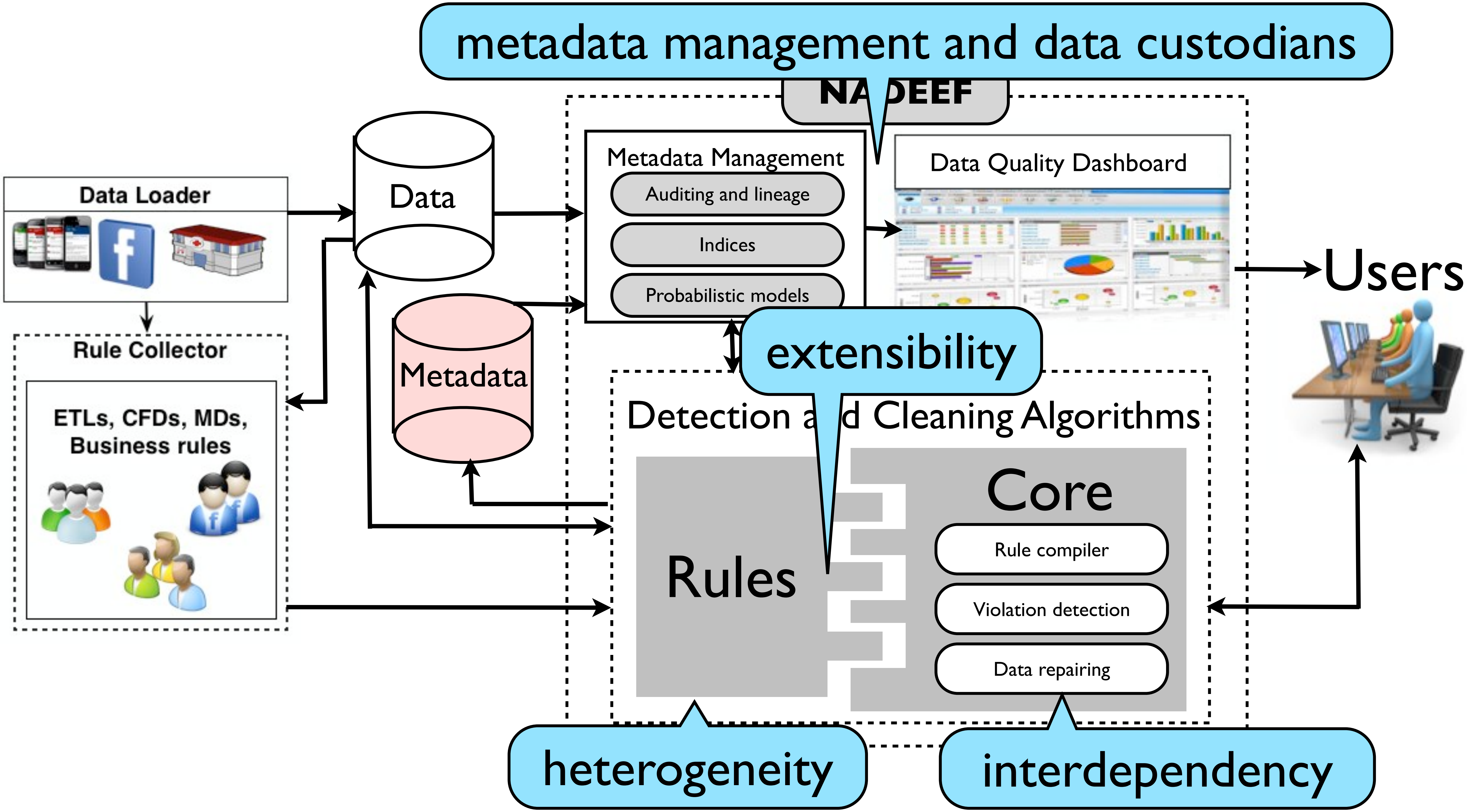
Violation detection

Data repairing

NADEEF Architecture



NADEEF Architecture



metadata management and data custodians

NADEEF

Data Loader

Data

Metadata Management

Auditing and lineage

Indices

Probabilistic models

Data Quality Dashboard

Users

Rule Collector

ETLs, CFDs, MDs, Business rules

Metadata

extensibility

Detection and Cleaning Algorithms

Rules

Core

Rule compiler

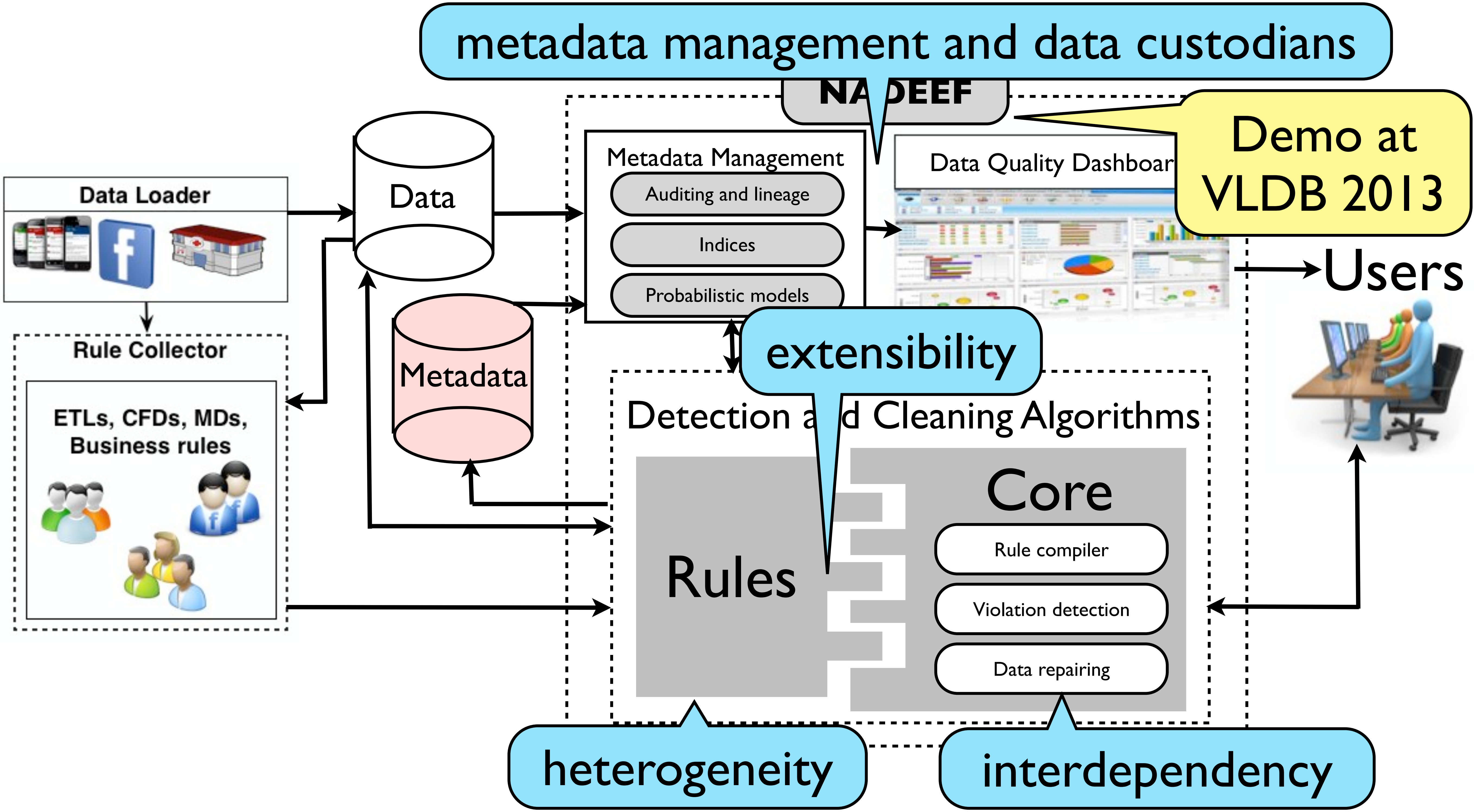
Violation detection

Data repairing

heterogeneity

interdependency

NADEEF Architecture



metadata management and data custodians

Demo at VLDB 2013

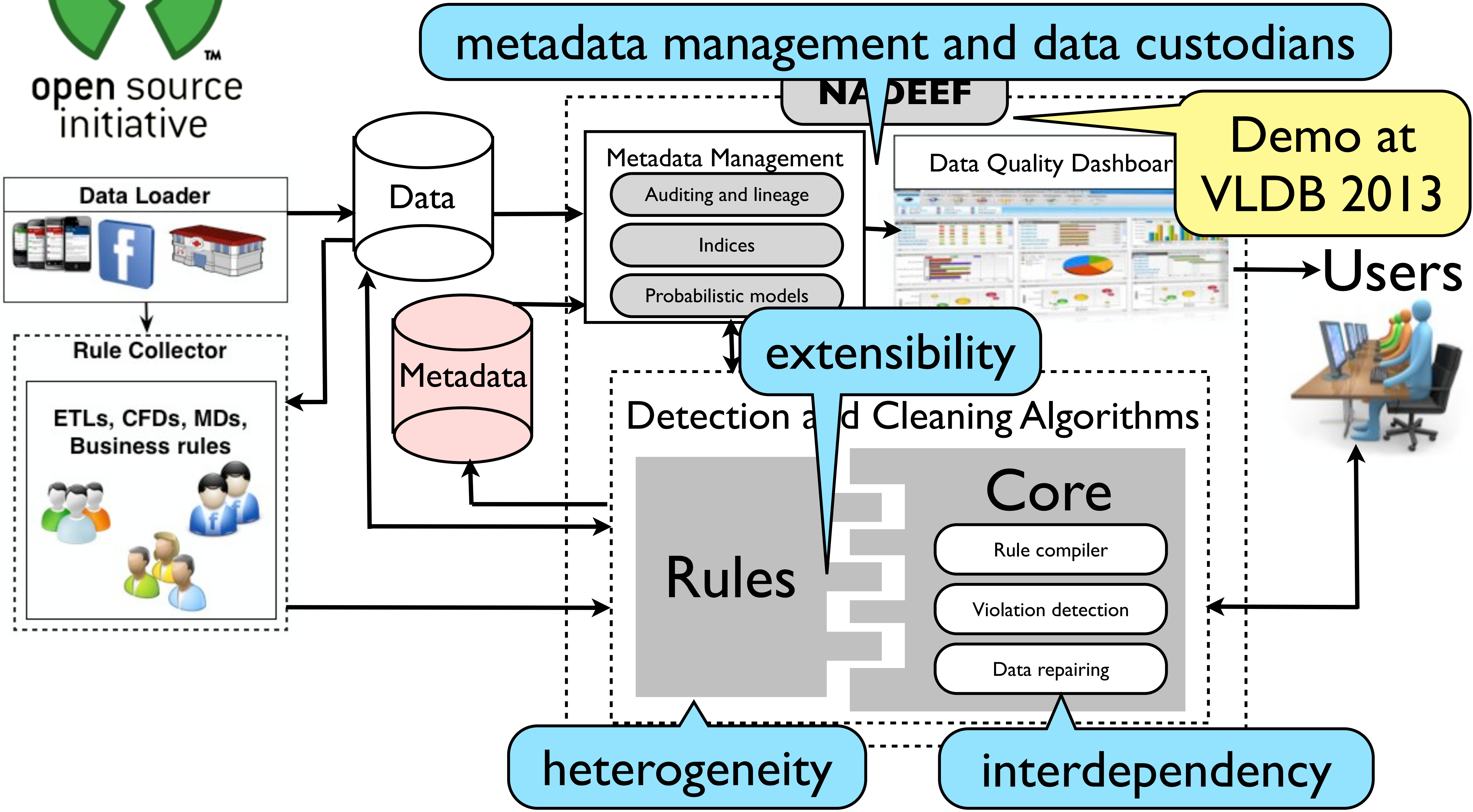
extensibility

heterogeneity

interdependency

open source initiative

NADEEF Architecture



metadata management and data custodians

NADEEF

Demo at VLDB 2013

Data Loader

Data

Metadata Management

Auditing and lineage

Indices

Probabilistic models

Data Quality Dashboar

Users

Rule Collector

ETLs, CFDs, MDs, Business rules

Metadata

extensibility

Detection and Cleaning Algorithms

Core

Rules

Rule compiler

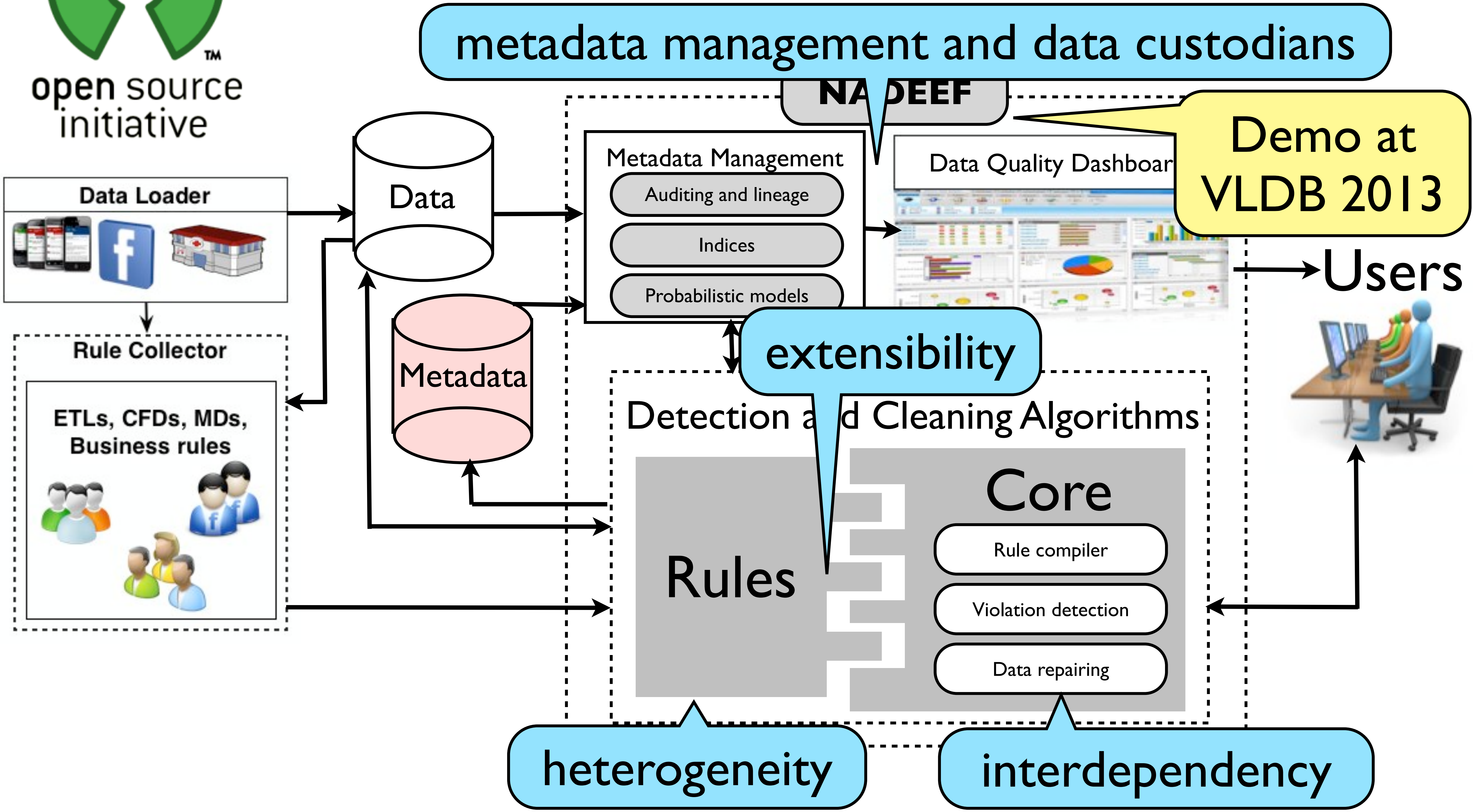
Violation detection

Data repairing

heterogeneity

interdependency

NADEEF Architecture



metadata management and data custodians

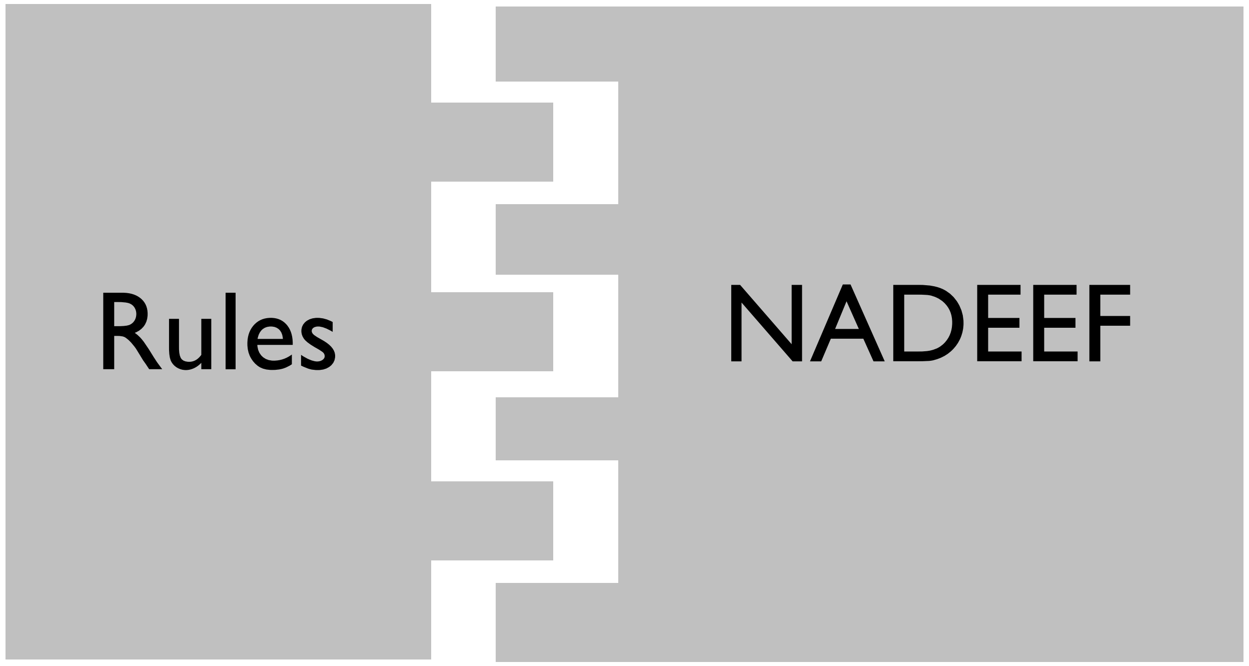
Demo at VLDB 2013

extensibility

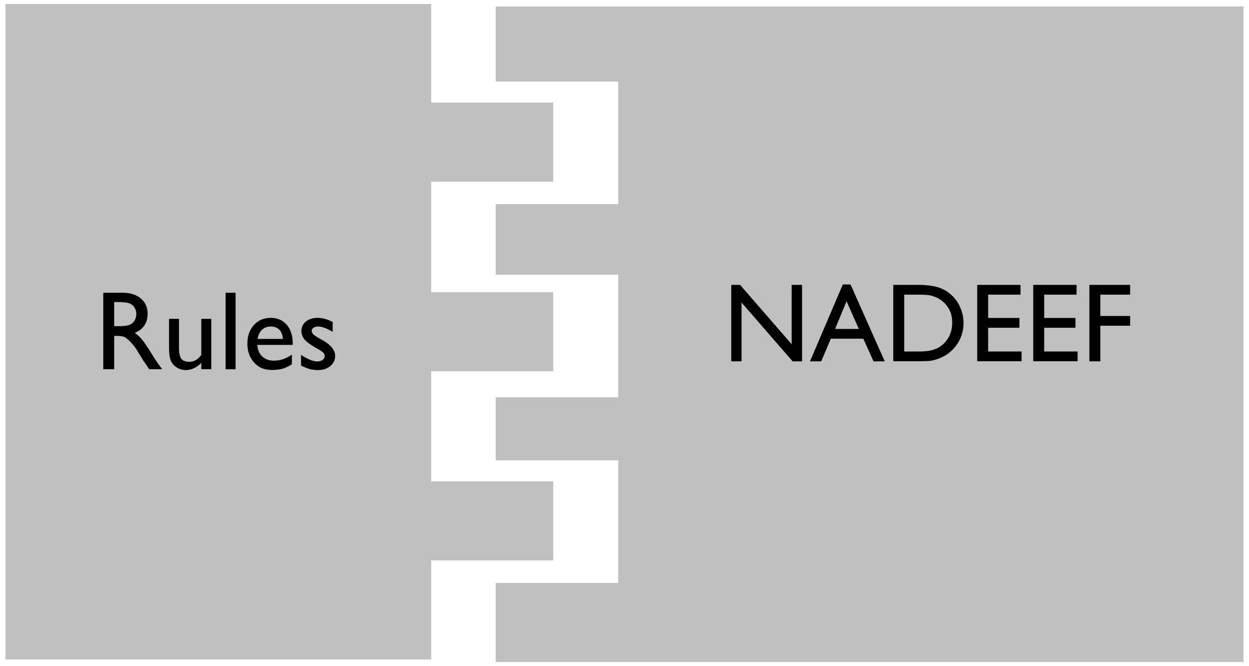
heterogeneity

interdependency

Programming Interface



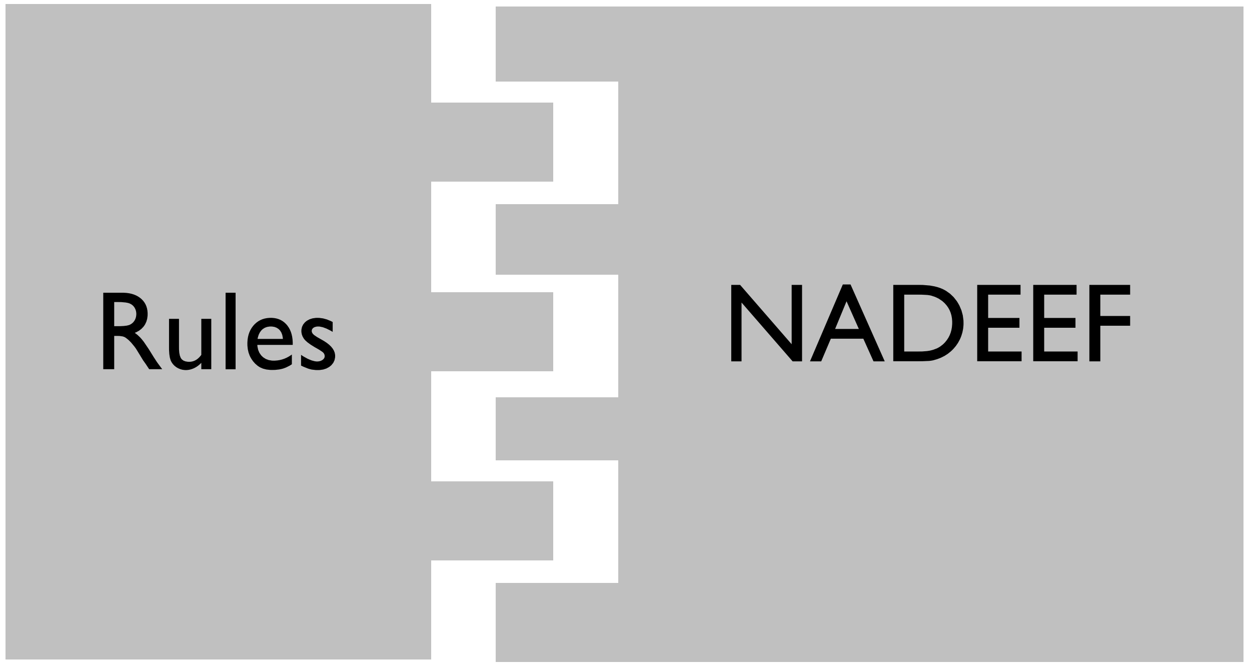
Programming Interface



Rule

static semantics	dynamic semantics
vio(tuple t)	fix(violation V)
vio(tuple t_1, tuple t_2)	

Programming Interface



Rule

static semantics	dynamic semantics
vio(tuple t)	fix(violation V)
vio(tuple t_1, tuple t_2)	

A simple and declarative interface

Sample Rules

(**tran**) If a customer's CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands.

Sample Rules

(**tran**) If a customer's CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands.

```
Class Rule1 {

set⟨cell⟩ vio(tuple t) { /*s in table tran */
    if (t[CC] = 31 ∧ !(t[country] = Netherlands ∨ t[country] = Holland))
        return { t[CC, country]; }
    return ∅;
}

set⟨Expression⟩ fix (set⟨cell⟩ V) {
    return { V.t[country] ← Netherlands; }
}

}
```

Sample Rules

(**tran**) If a customer's CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands.

```
Class Rule1 {

set⟨cell⟩ vio(tuple t) { /*s in table tran */
    if (t[CC] = 31 ∧ !(t[country] = Netherlands ∨ t[country] = Holland))
        return { t[CC, country]; }
    return ∅;
}

set⟨Expression⟩ fix (set⟨cell⟩ V) {
    return { V.t[country] ← Netherlands; }
}

}
```

static semantics: what is wrong

Sample Rules

(**tran**) If a customer's CC is 31, but his/her country is neither Netherlands nor Holland, update the country to Netherlands.

```
Class Rule1 {

set⟨cell⟩ vio(tuple t) { /*s in table tran */
    if (t[CC] = 31 ∧ !(t[country] = Netherlands ∨ t[country] = Holland))
        return { t[CC, country]; }
    return ∅;
}
```

static semantics: what is wrong

```
set⟨Expression⟩ fix (set⟨cell⟩ V) {
    return { V.t[country] ← Netherlands; }
}
}
```

dynamic semantics: possible ways to repair

Sample Rules

(**tran**) If two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour, these two purchases might be a fraud.

Sample Rules

(**tran**) If two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour, these two purchases might be a fraud.

Class Rule4 {

```
set⟨cell⟩ vio(tuple $t_1$, tuple $t_2$) { /* $t_1$, $t_2$ in table tran */
  if ($t_1$[name] ≈ $t_2$[name] ∧ $t_1$[tel] = $t_2$[tel] ∧ $t_1$[where] = Netherlands
        ∧ $t_2$[where] = US ∧ | $t_1$[when] - $t_2$[when] | <= 1 )
     return { $t_1$[name, tel, where, when]; $t_2$[name, tel, where, when]; }
  return ∅;
}
```

}

Sample Rules

(**tran**) If two purchases of the same person happened in the Netherlands and the US (East Coast) within 1 hour, these two purchases might be a fraud.

Class Rule4 {

```
set⟨cell⟩ vio(tuple $t_1$, tuple $t_2$) { /* $t_1$, $t_2$ in table tran */
  if ($t_1$[name] $\approx$ $t_2$[name] $\wedge$ $t_1$[tel] = $t_2$[tel] $\wedge$ $t_1$[where] = Netherlands
        $\wedge$ $t_2$[where] = US $\wedge$ | $t_1$[when] - $t_2$[when] | <= 1 )
     return { $t_1$[name, tel, where, when]; $t_2$[name, tel, where, when]; }
  return $\emptyset$;
}
```

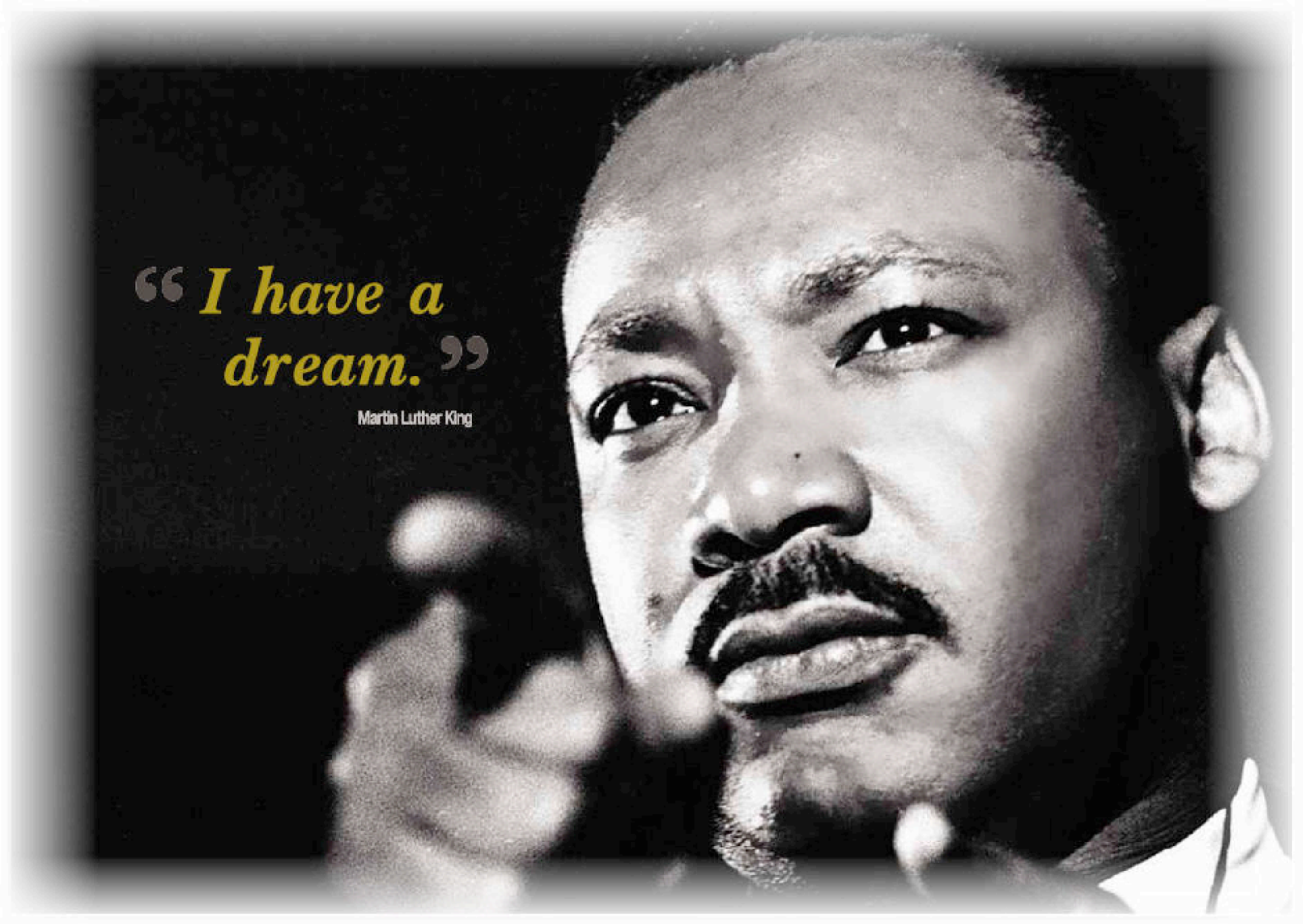
static semantics: what is wrong

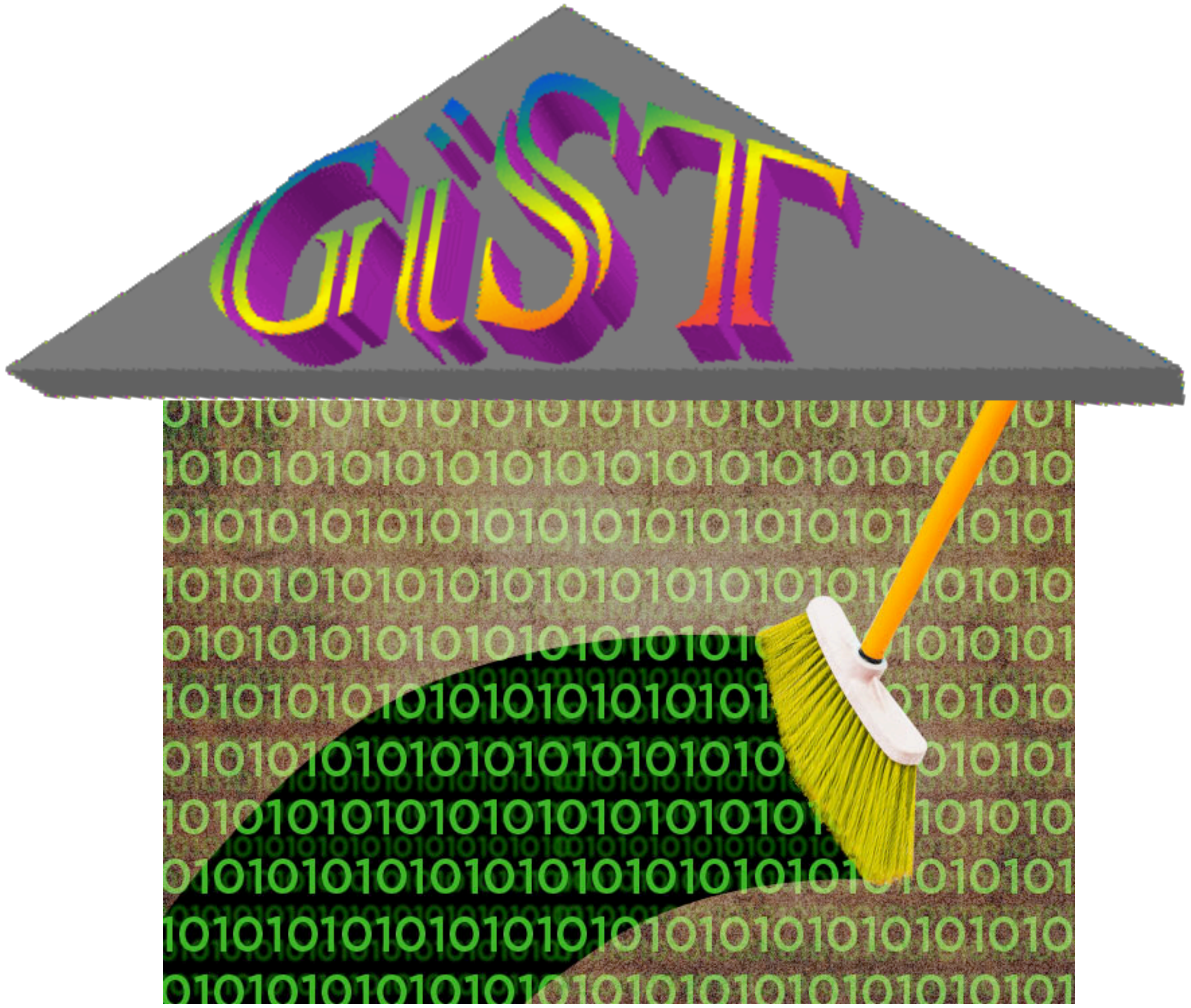
}

NADEEF Extensibility

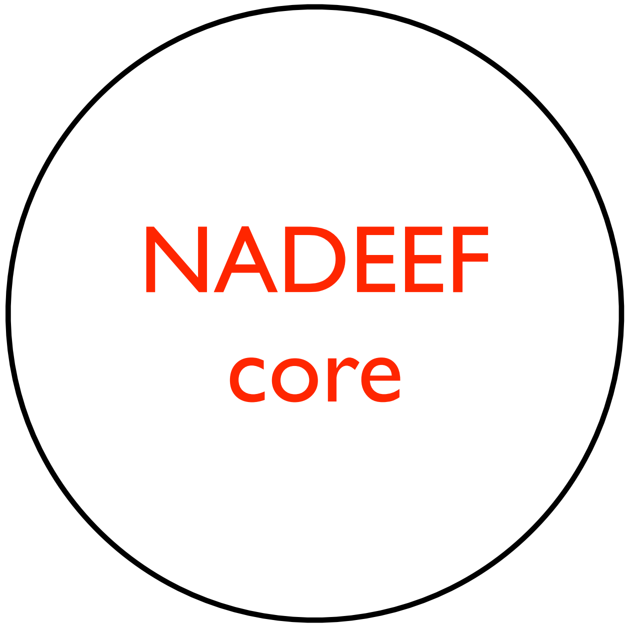
NADEEF Extensibility

NADEEF Extensibility

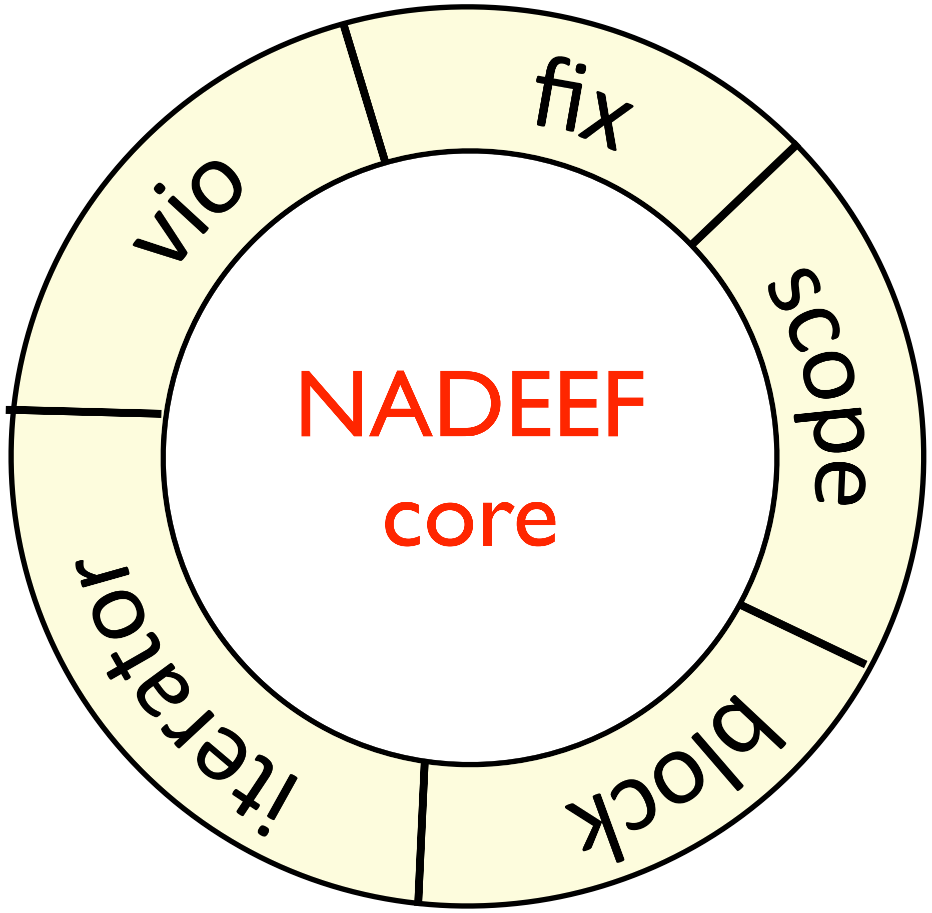




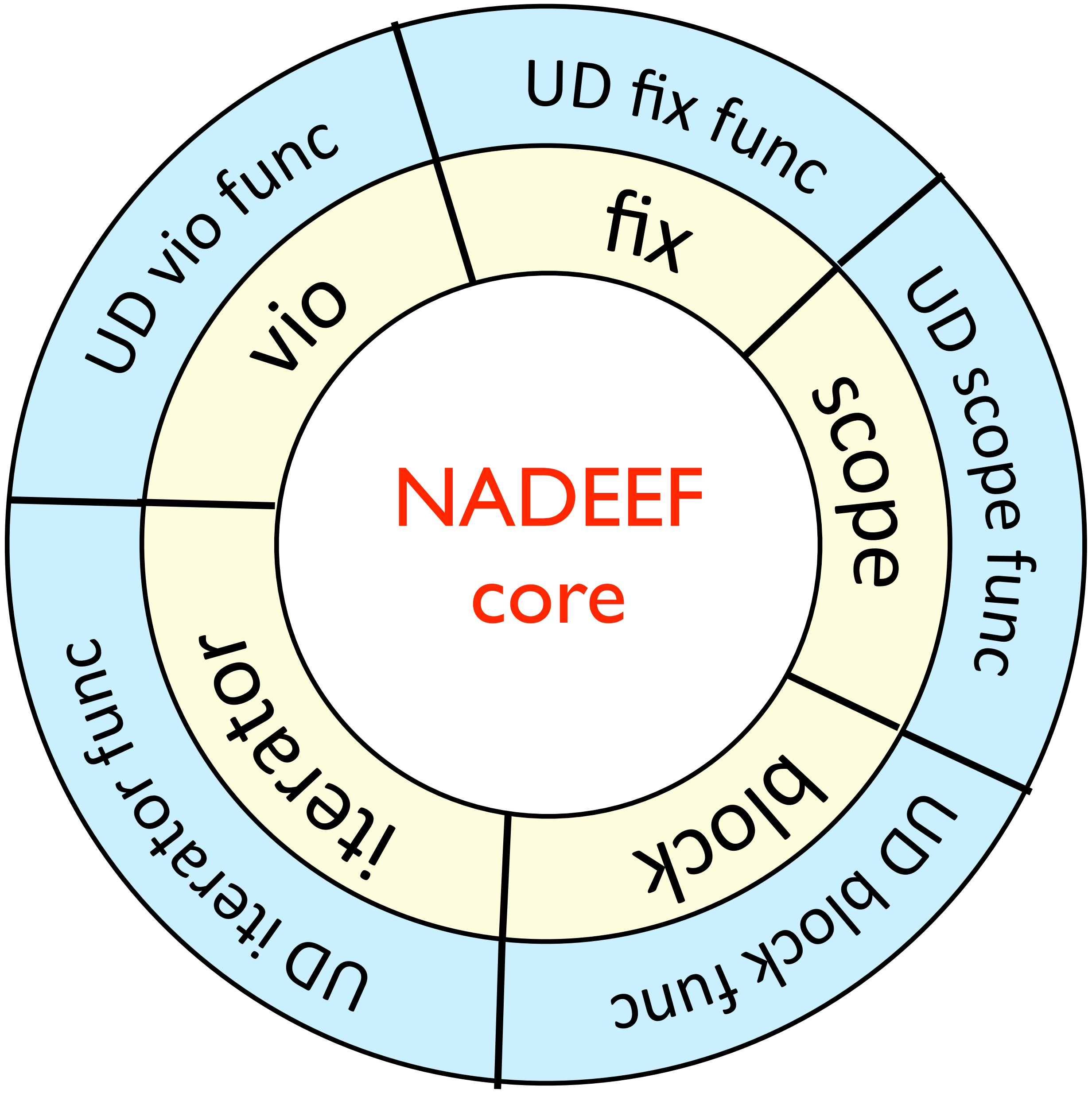
NADEEF Extensibility



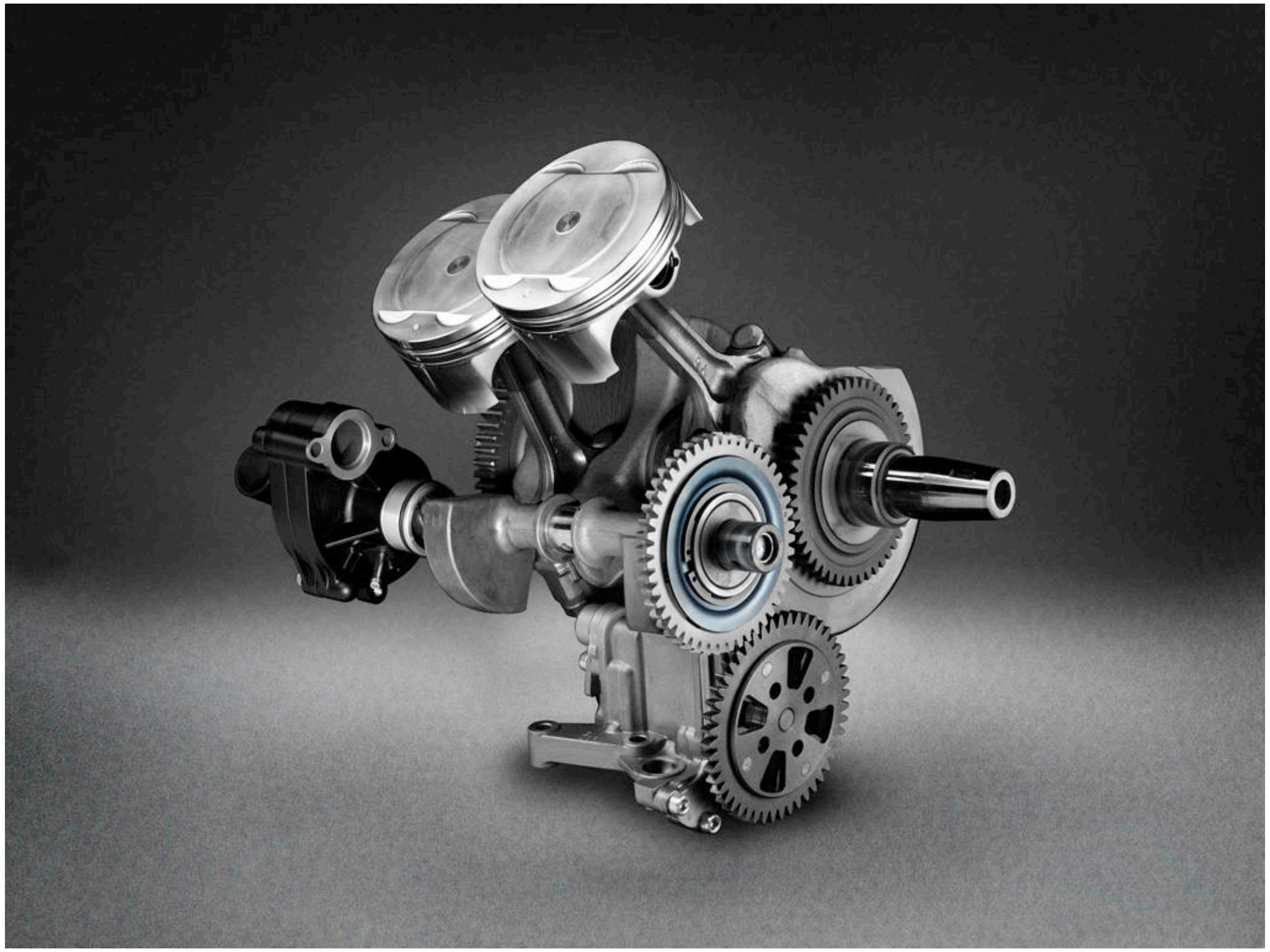
NADEEF Extensibility



NADEEF Extensibility



Inside NADEEF



Inside NADEEF

Rule 1
Rule 2
Rule 3
Rule 4

tran

name	street	city	CC	country	phn	when	where
David	Holywell	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
Paul	Ratcliffe	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
David	Holywell	Oxford	44	Netherlands	66700541	6am 6/05/2012	US
Paul	Market	Amsterdam	31	UK	55384922	9am 6/02/2012	Netherlands

bank

name	street	city	CC	country	tel	gd
David	Holywell	Oxford	44	UK	66700543	M
Paul	Ratcliffe	Oxford	44	UK	44944631	M

Inside NADEEF

Rule 1
Rule 2
Rule 3
Rule 4

tran

name	street	city	CC	country	phn	when	where
David	Holywell	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
Paul	Ratcliffe	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
David	Holywell	Oxford	44	UK	66700543	6am 6/05/2012	US
Paul	Market	Amsterdam	31	Netherlands	55384922	9am 6/02/2012	Netherlands

bank

name	street	city	CC	country	tel	gd
David	Holywell	Oxford	44	UK	66700543	M
Paul	Ratcliffe	Oxford	44	UK	44944631	M

Inside NADEEF

Rule 1
Rule 2
Rule 3
Rule 4

tran

name	street	city	CC	country	phn	when	where
David	Holywell	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
Paul	Ratcliffe	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
David	Holywell	Oxford	44	UK	66700543	6am 6/05/2012	US
Paul	Market	Amsterdam	31	Netherlands	55384922	9am 6/02/2012	Netherlands

bank

name	street	city	CC	country	tel	gd
David	Holywell	Oxford	44	UK	66700543	M
Paul	Ratcliffe	Oxford	44	UK	44944631	M

Inside NADEEF

Rule 1
Rule 2
Rule 3
Rule 4

tran

name	street	city	CC	country	phn	when	where
David	Holywell	Oxford	44	UK	66700543	1pm 6/05/2012	Netherlands
Paul	Ratcliffe	Oxford	44	UK	44944631	11am 2/12/2011	Netherlands
David	Holywell	Oxford	44	UK	66700543	6am 6/05/2012	US
Paul	Market	Amsterdam	31	Netherlands	55384922	9am 6/02/2012	Netherlands

bank

name	street	city	CC	country	tel	gd
David	Holywell	Oxford	44	UK	66700543	M
Paul	Ratcliffe	Oxford	44	UK	44944631	M

Violation Detection

Violation Detection

- Brute force approach (black-boxes)

	CC	country	...
r1	44	UK	...
r2	44	UK	...
r3	44	Netherlands	...
r4	31	UK	...

Violations:
(r1, r3), (r2, r3)

Violation Detection

- Brute force approach (black-boxes)
- Optimized approach (white-boxes, e.g., CC->country)

	CC	country	...
r1	44	UK	...
r2	44	UK	...
r3	44	Netherlands	...
r4	31	UK	...

Violations:
(r1, r3), (r2, r3)

Violation Detection

- Brute force approach (black-boxes)
- Optimized approach (white-boxes, e.g., CC->country)

	CC	country	...
r1	44	UK	...
r2	44	UK	...
r3	44	Netherlands	...
r4	31	UK	...

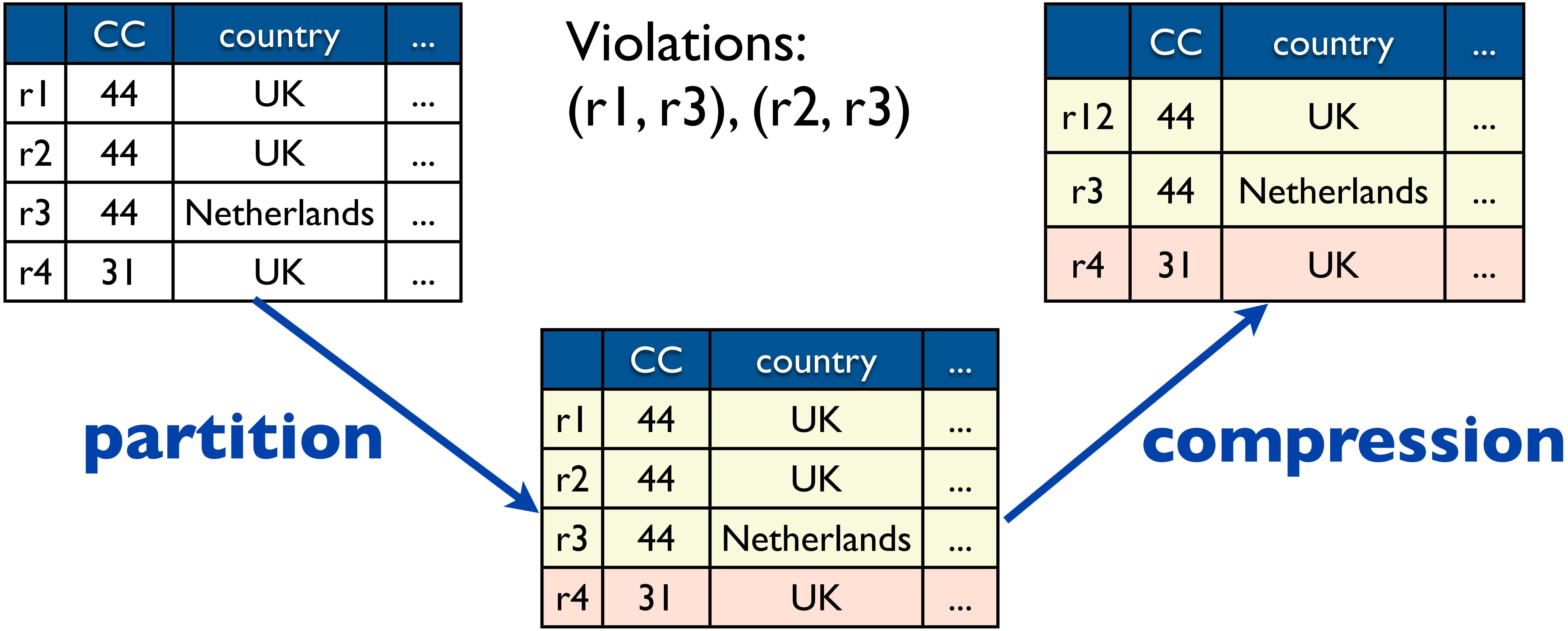
Violations:
(r1, r3), (r2, r3)

partition

	CC	country	...
r1	44	UK	...
r2	44	UK	...
r3	44	Netherlands	...
r4	31	UK	...

Violation Detection

- Brute force approach (black-boxes)
- Optimized approach (white-boxes, e.g., CC->country)



	CC	country	...
r1	44	UK	...
r2	44	UK	...
r3	44	Netherlands	...
r4	31	UK	...

Violations:
(r1, r3), (r2, r3)

partition

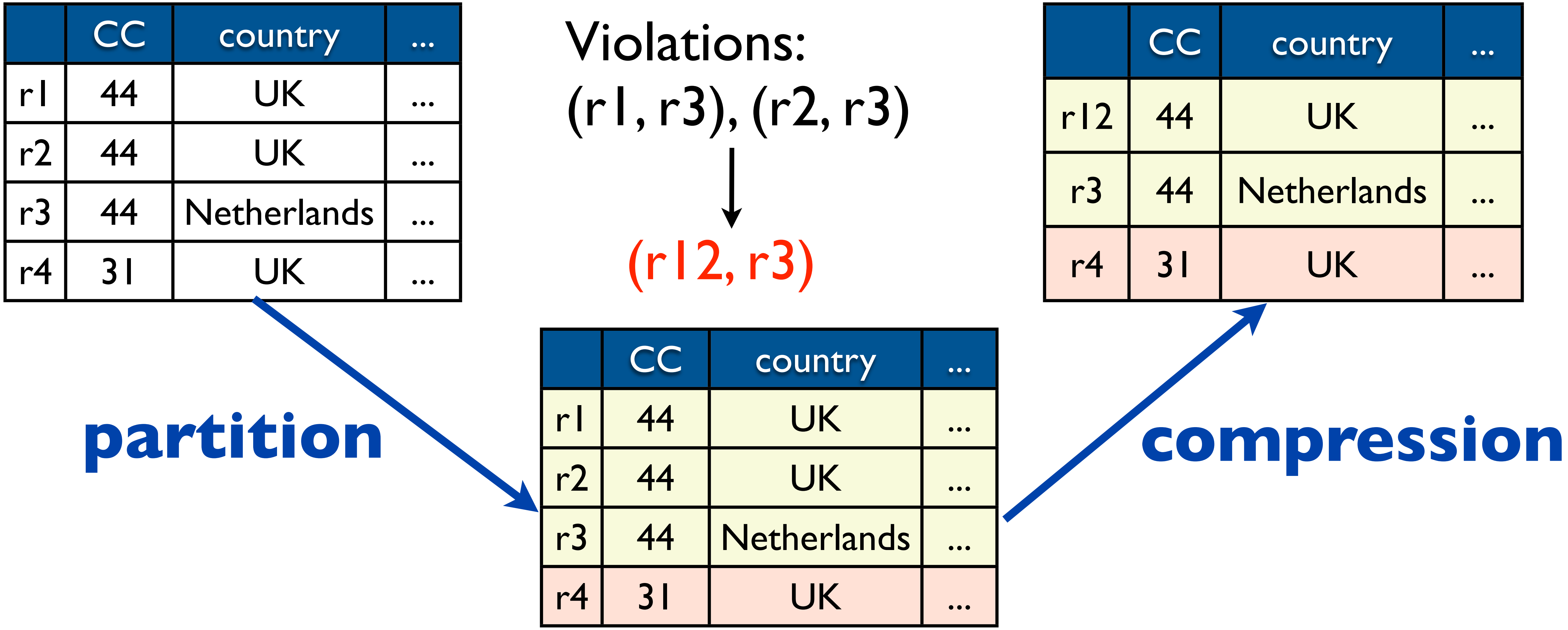
	CC	country	...
r1	44	UK	...
r2	44	UK	...
r3	44	Netherlands	...
r4	31	UK	...

compression

	CC	country	...
r12	44	UK	...
r3	44	Netherlands	...
r4	31	UK	...

Violation Detection

- Brute force approach (black-boxes)
- Optimized approach (white-boxes, e.g., CC->country)



	CC	country	...
r1	44	UK	...
r2	44	UK	...
r3	44	Netherlands	...
r4	31	UK	...

Violations:
(r1, r3), (r2, r3)

(r12, r3)

	CC	country	...
r1	44	UK	...
r2	44	UK	...
r3	44	Netherlands	...
r4	31	UK	...

	CC	country	...
r12	44	UK	...
r3	44	Netherlands	...
r4	31	UK	...

Data Repairing

Holistic Data Cleaning

data cleaning

rule specification

Holistic Data Cleaning

Violations
V1: {r4[CC, country]}
V2: {t1[name, street, city, tel], r3[name, street, city, phn]}
V3: {r1[CC,country], r3[CC, country]}
V4: {r2[CC,country], r3[CC, country]}
V5: {r1[name, tel, where, when], r3[name, tel, where, when]}

data cleaning

rule specification

Holistic Data Cleaning

Violations
V1: {r4[CC, country]}
V2: {t1[name, street, city, tel], r3[name, street, city, phn]}
V3: {r1[CC,country], r3[CC, country]}
V4: {r2[CC,country], r3[CC, country]}
V5: {r1[name, tel, where, when], r3[name, tel, where, when]}

data cleaning

rule specification

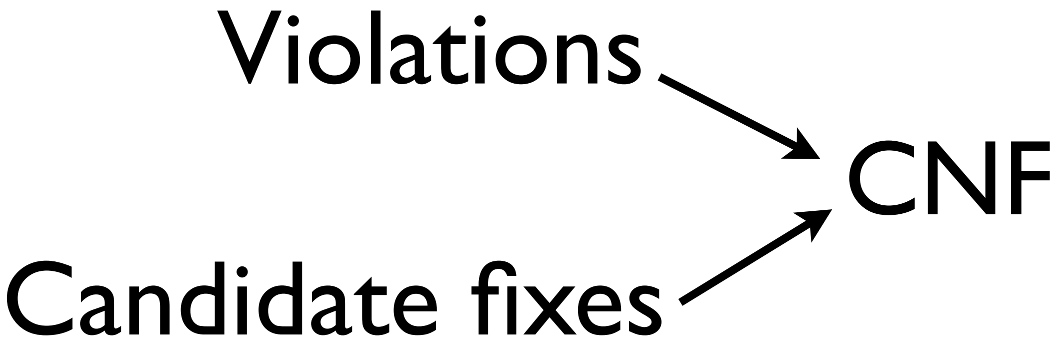
Candidate fixes
F1: r4[country]←Netherlands
F2: r3[phn] ← t1[tel]
F3: r1[country] ← r3[country]
F4: r3[country] ← r1[country]
F5: r2[country] ← r3[country]
F6: r3[country] ← r2[country]

A Variable-Weighted Max-SAT

Violations

Candidate fixes

A Variable-Weighted Max-SAT

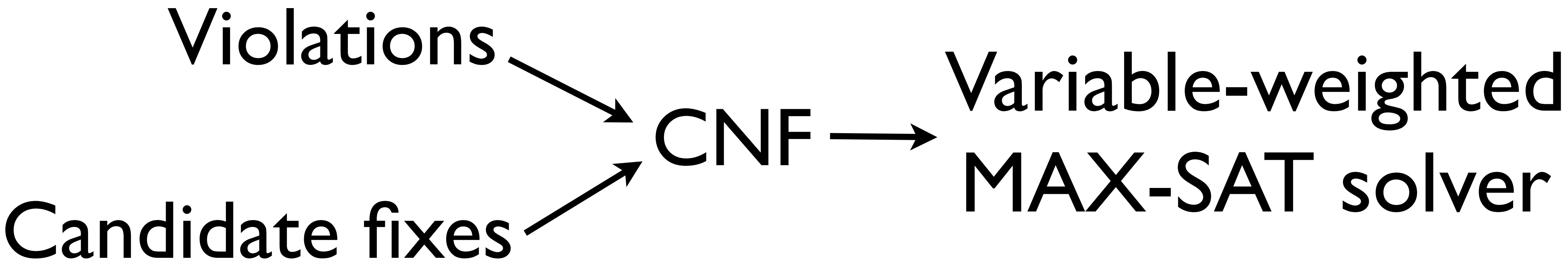


Violations

CNF

Candidate fixes

A Variable-Weighted Max-SAT



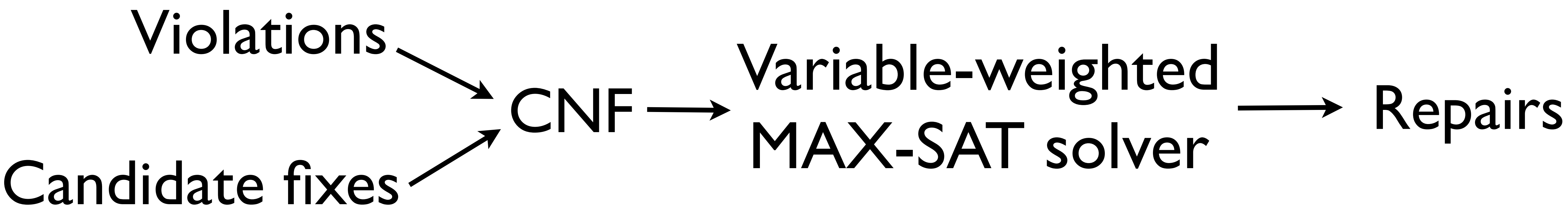
Violations

Candidate fixes

CNF

Variable-weighted MAX-SAT solver

A Variable-Weighted Max-SAT



Violations

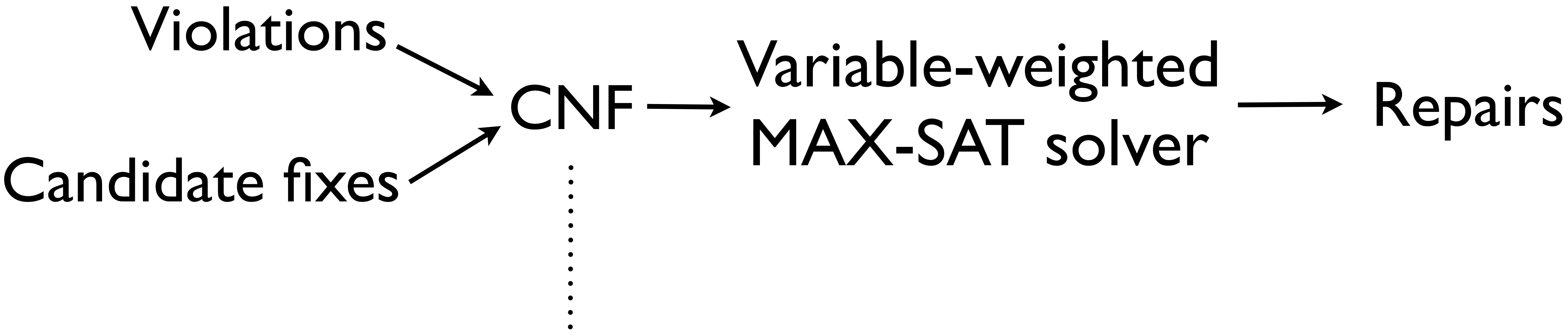
Candidate fixes

CNF

Variable-weighted MAX-SAT solver

Repairs

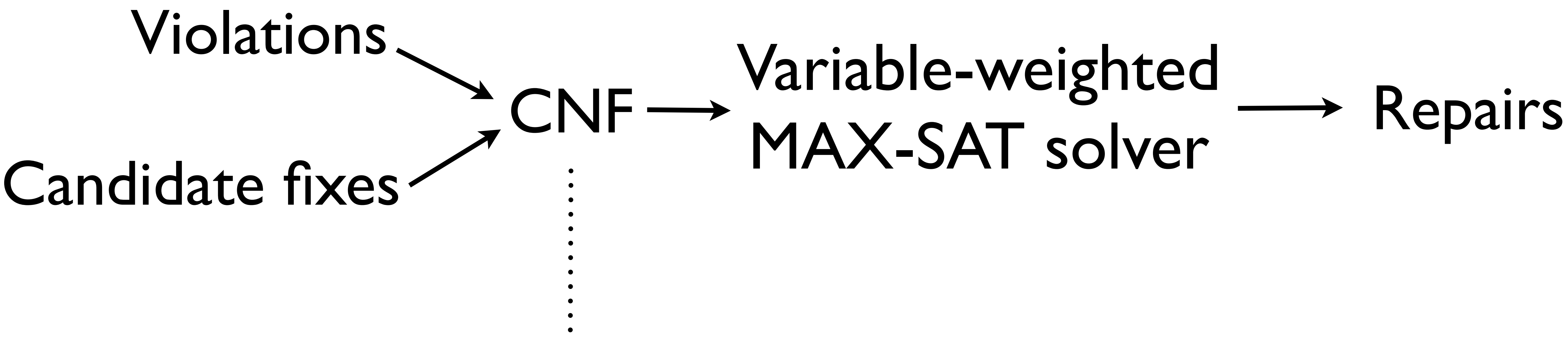
A Variable-Weighted Max-SAT



variable

$x_{r_4[\text{country}]}^{Netherlands}$

A Variable-Weighted Max-SAT



Violations

Candidate fixes

CNF

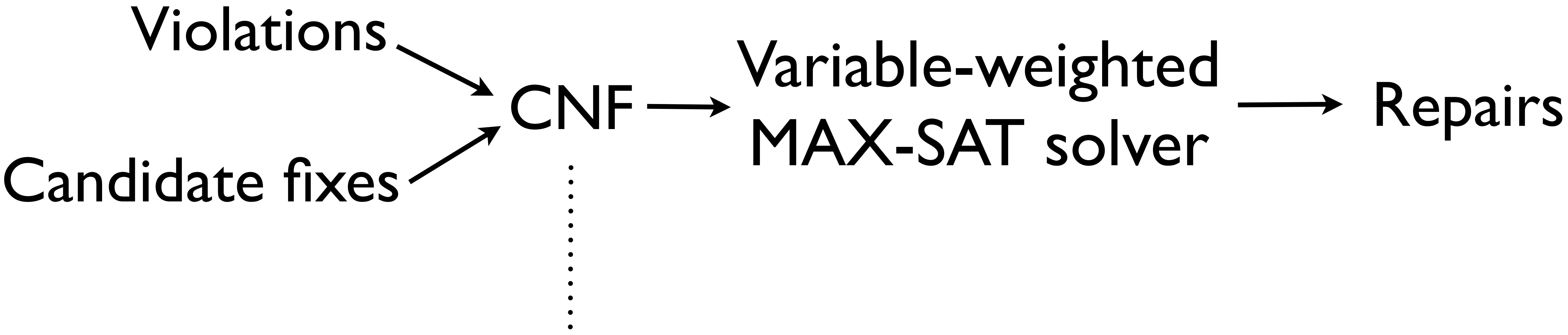
Variable-weighted MAX-SAT solver

Repairs

variable $x^{Netherlands}_{r_4[\text{country}]}$

inclusive assignment $(x^{UK}_{r_4[\text{country}]} \vee x^{Netherlands}_{r_4[\text{country}]})$

A Variable-Weighted Max-SAT



Violations

Candidate fixes

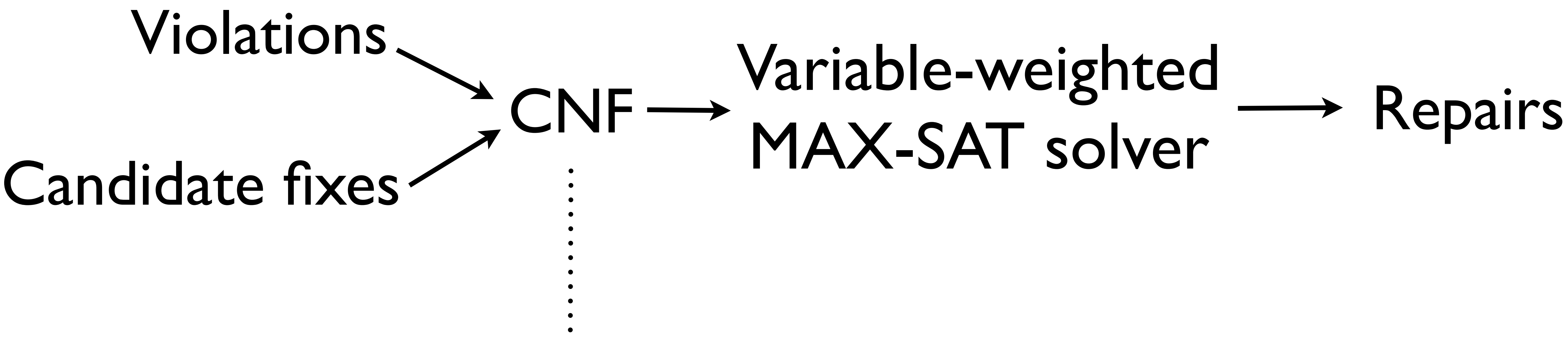
CNF

Variable-weighted MAX-SAT solver

Repairs

variable	$x^{Netherlands}_{r_4[\text{country}]}$
inclusive assignment	$(x^{UK}_{r_4[\text{country}]} \vee x^{Netherlands}_{r_4[\text{country}]})$
exclusive assignment	$(\neg x^{UK}_{r_4[\text{country}]} \vee \neg x^{Netherlands}_{r_4[\text{country}]})$

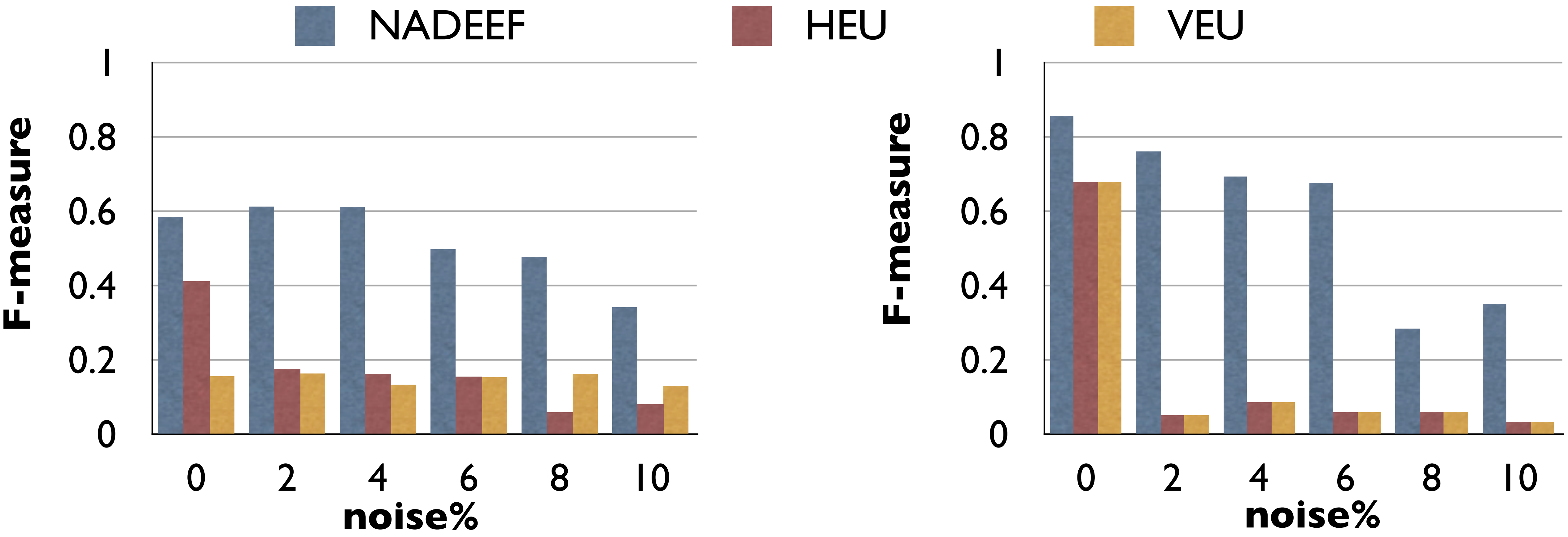
A Variable-Weighted Max-SAT



variable	$x^{Netherlands}_{r_4[\text{country}]}$
inclusive assignment	$(x^{UK}_{r_4[\text{country}]} \vee x^{Netherlands}_{r_4[\text{country}]})$
exclusive assignment	$(\neg x^{UK}_{r_4[\text{country}]} \vee \neg x^{Netherlands}_{r_4[\text{country}]})$
avoid violations	$(\neg x^{31}_{r_4[\text{CC}]} \vee \neg x^{UK}_{r_4[\text{country}]})$

Experimental Study

Effectiveness



(a) Hospital dataset
(100K, 9 attributes, 10 rules)

(b) Bus dataset
(160K, 16 attributes, 11 rules)

Conclusion

- A generalized programming interface (heterogeneity)
- Holistic data cleaning (interdependency)
- An extensible system (extensibility)